Proxy-FDA: Proxy-based Feature Distribution Alignment for Fine-tuning Foundation Models without Forgetting

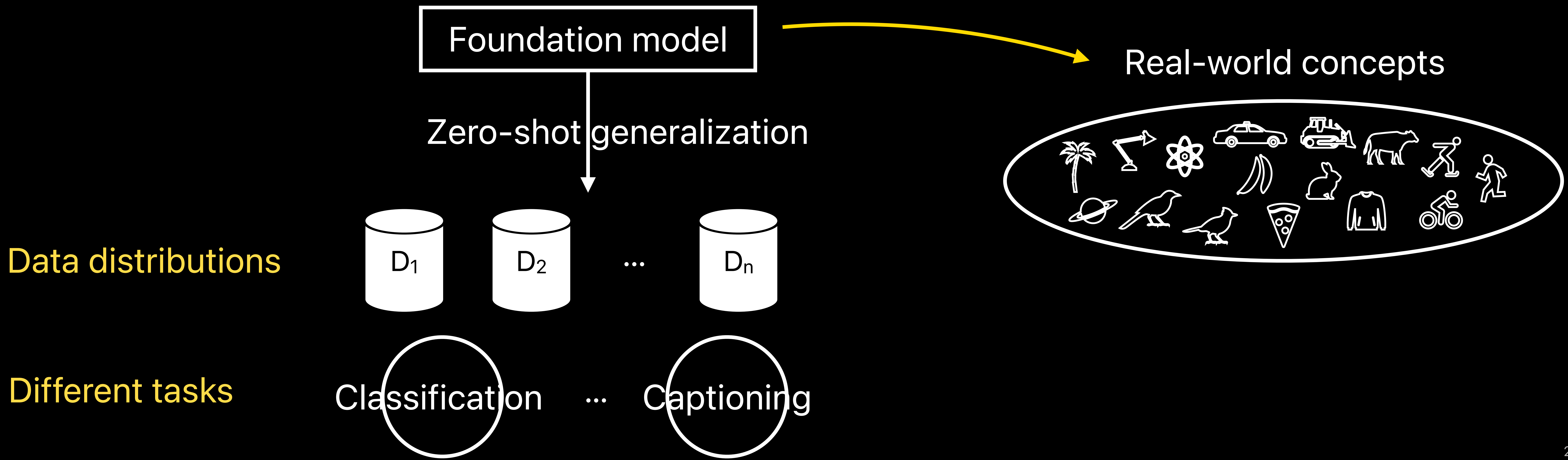
Chen Huang

ICML | Apple | July 14, 2025

Background

Foundation models show good zero-shot generalization

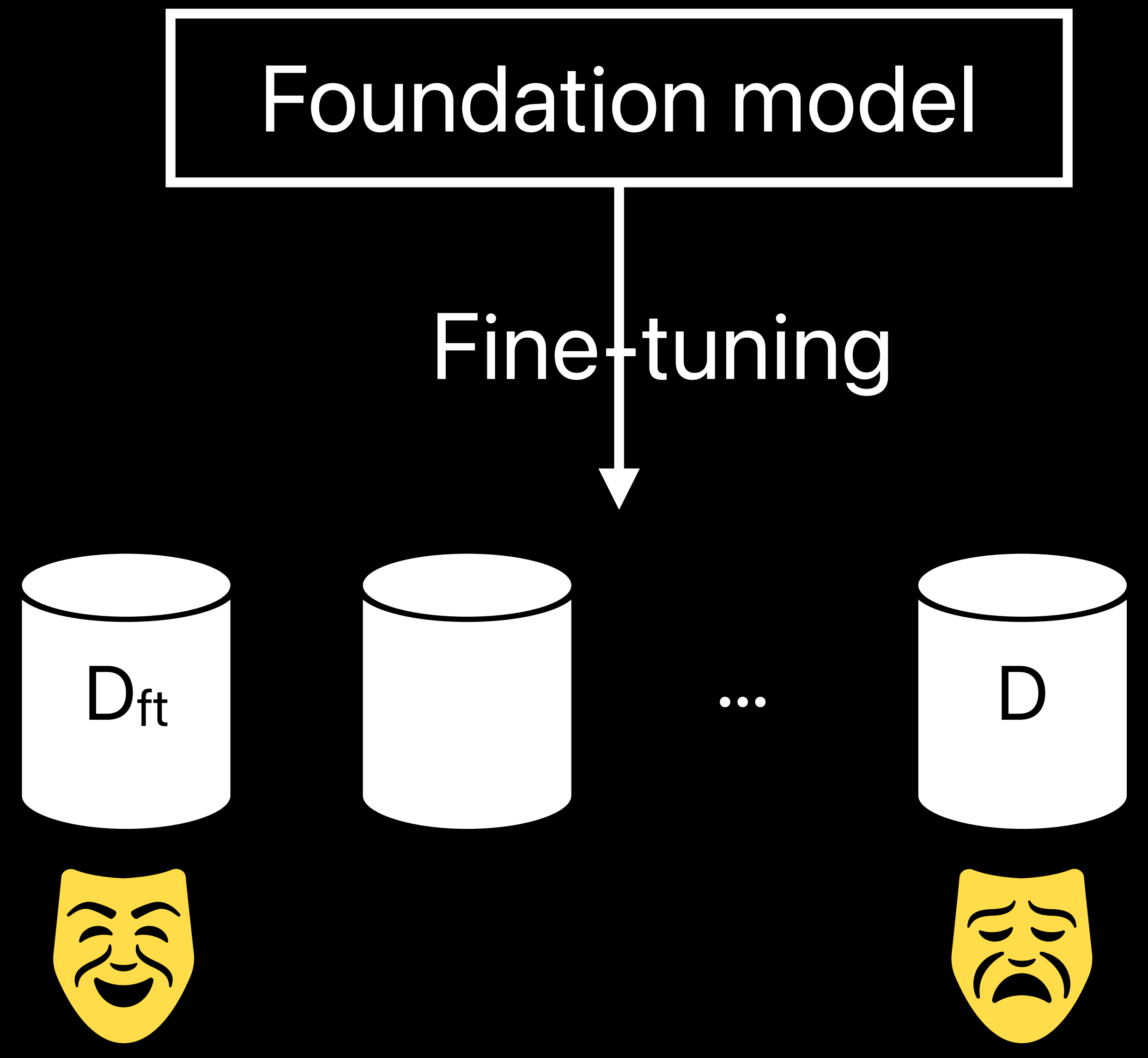
- Vision foundation models: CLIP, DINOv2, MAE…
- They encode rich knowledge on real-world concepts



Background

Fine-tuning for better task adaptation

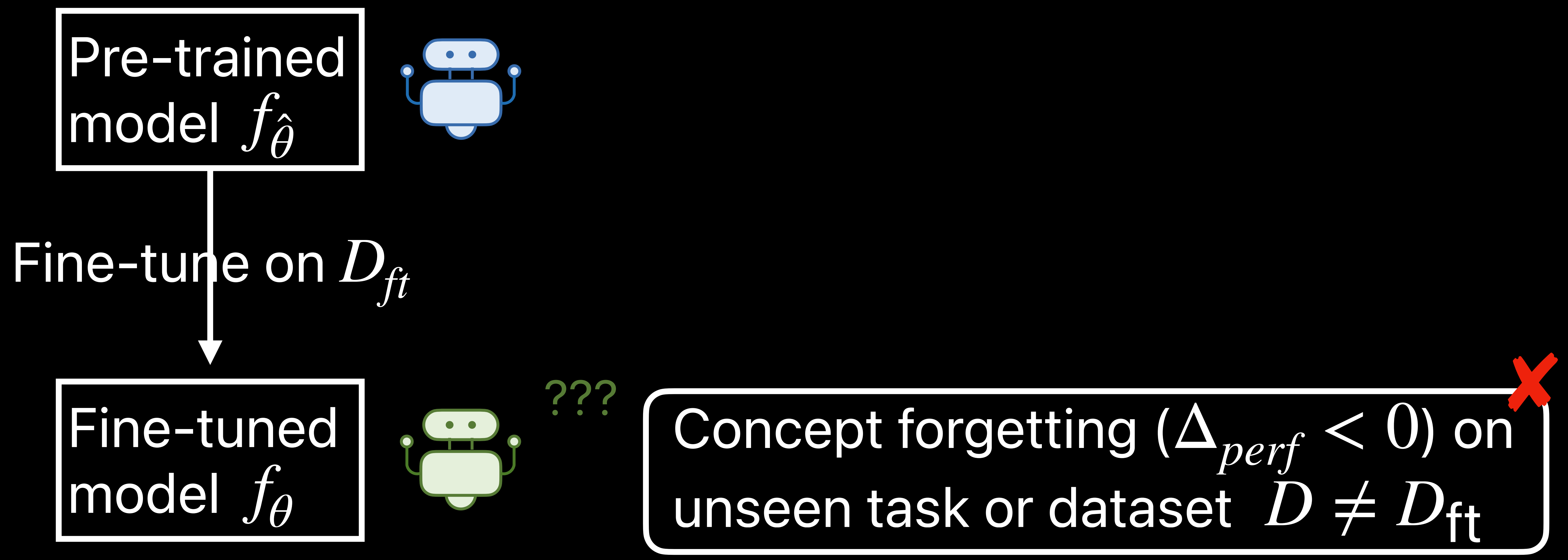
- End-to-end, linear probing
- Parameter-efficient fine-tuning: prompt tuning, adapter learning
- Frequent undesirable effect: concept forgetting



Background

Concept forgetting: after fine-tuning on D_{ft}, $\Delta_{perf} < 0$ between $f_{\hat{\theta}}$ and f_{θ} on a new dataset D

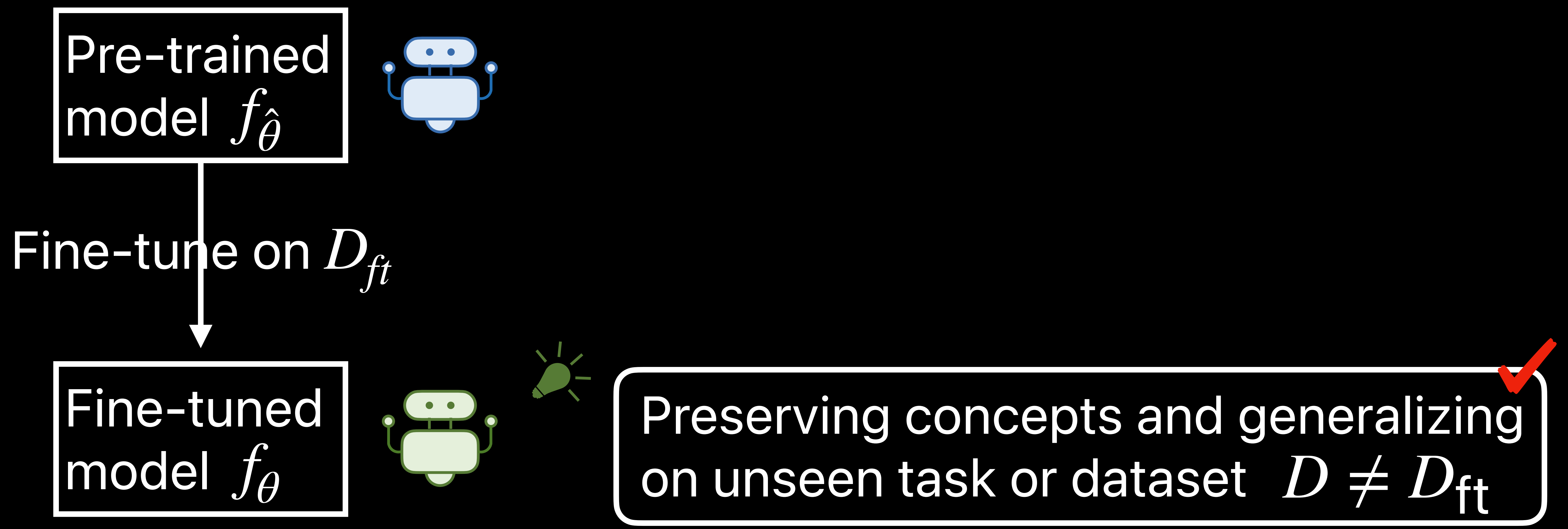
- Overfitting on downstream dataset D_{ft}



Goal

Robust fine-tuning of foundation models

- Preserve pre-trained knowledge, and generalize to new tasks
- Maintain good fine-tuning performance on the target task



Literature

Robust fine-tuning to mitigate concept forgetting

Two-stage tuning: linear probing + end-to-end

- LP-FT [ICLR 2022]

Ensemble models before and after fine-tuning

- WiSE-FT [CVPR 2022]

Literature

Robust fine-tuning to mitigate concept forgetting

Weight-space regularization

- L2SP [ICML 2018]: constrain the change in model weights before and after fine-tuning

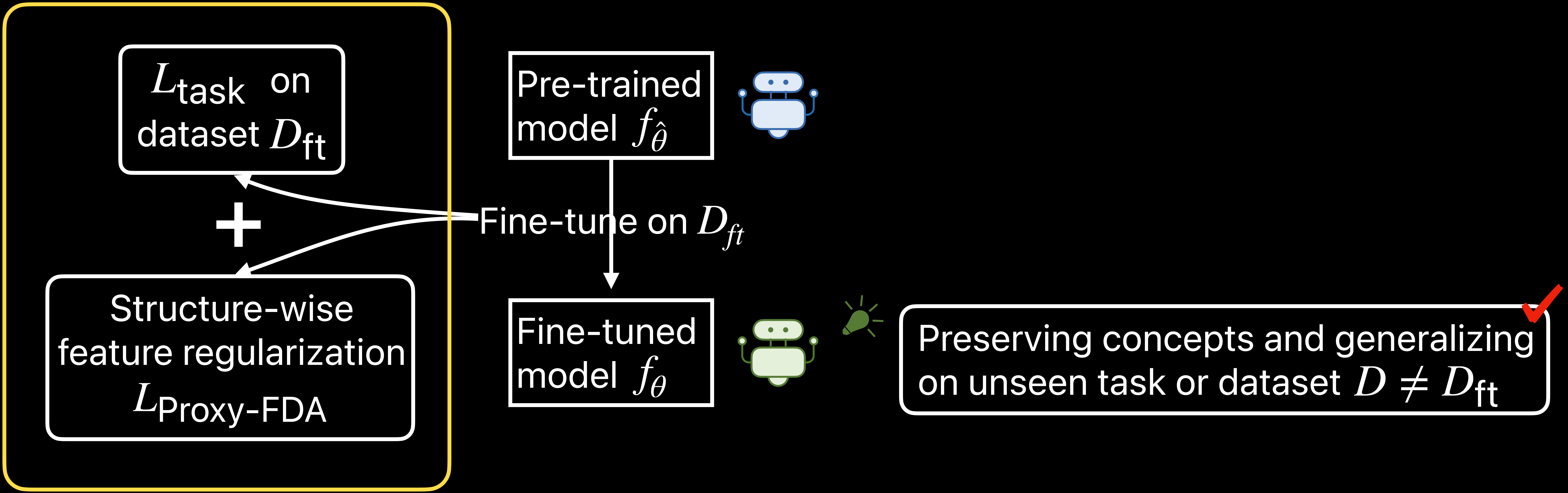
Feature-space regularization

- LDIFS [TMLR 2024]: match the pre-trained and fine-tuned features across samples
  - More promising: directly minimizes the change in input-output behavior of a model
  - Point-wise regularization is too strong
  - Lack explicit awareness of the feature neighborhood structures that encode rich knowledge too!

Idea

To better preserve concepts during fine-tuning

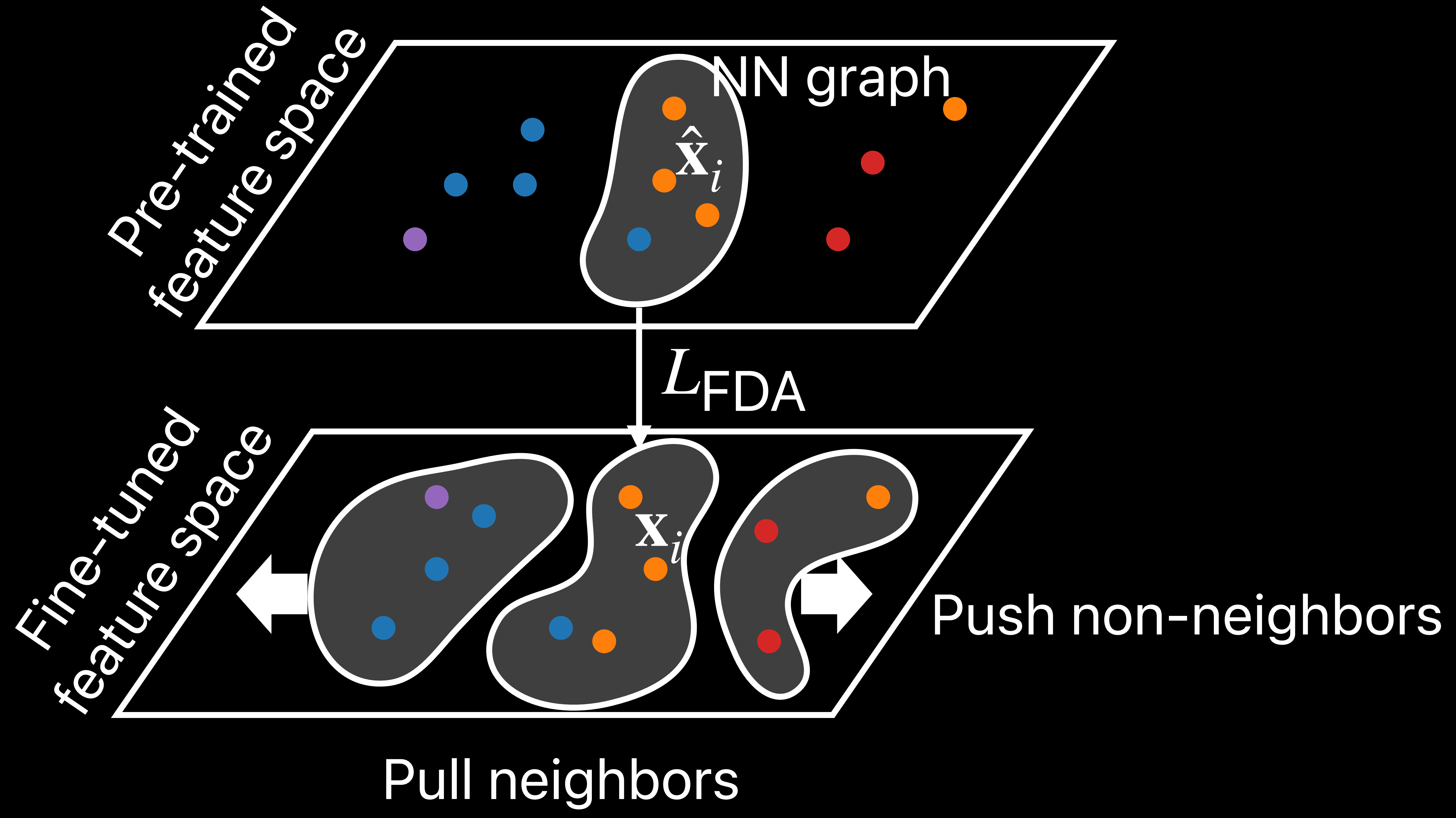
- Structure-wise feature regularization — Proxy-FDA



Method

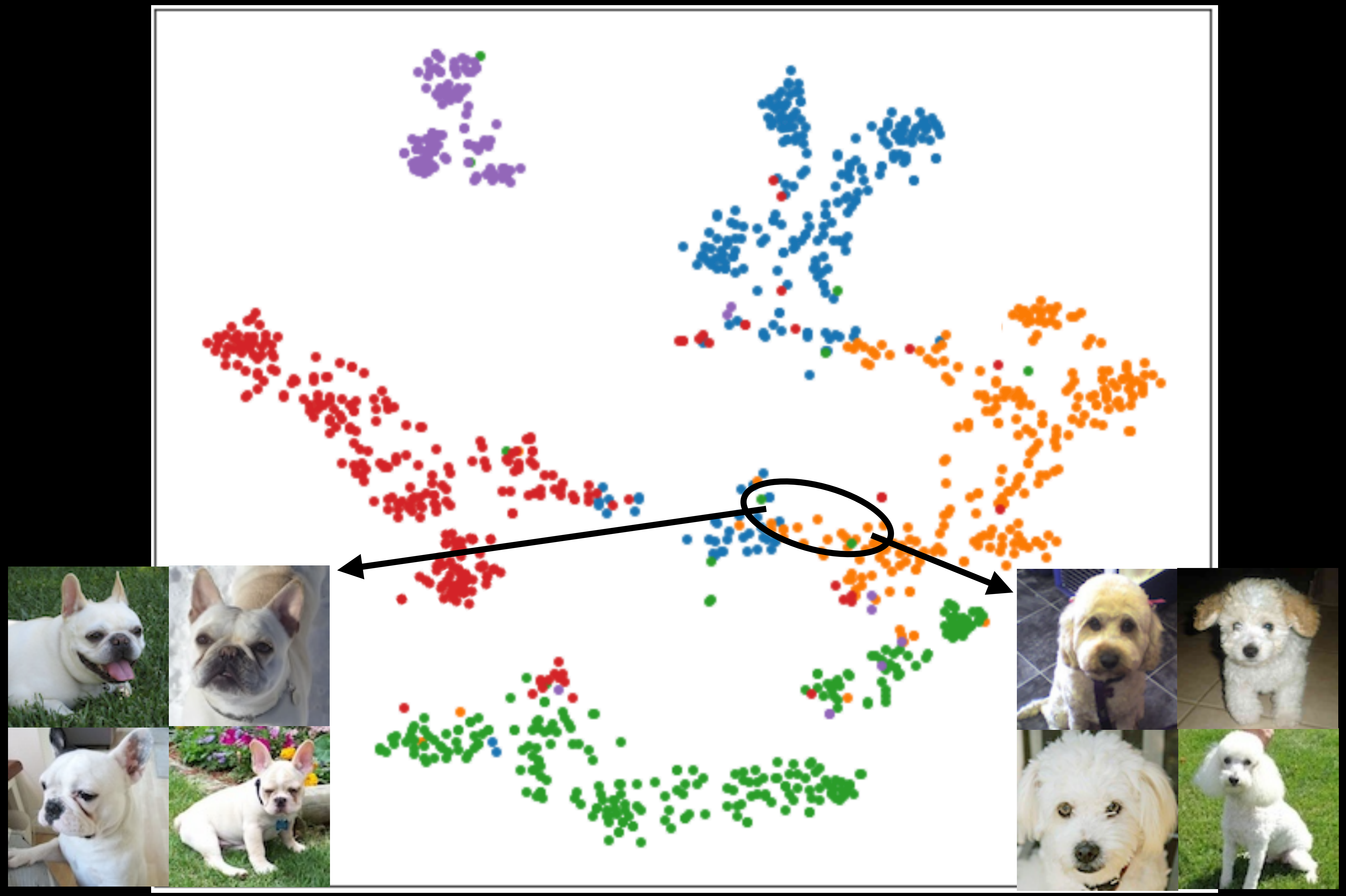
Feature Distribution Alignment (FDA)

- Align the structures of the pre-trained and fine-tuned feature distributions
- Structure in NN graph: neighbor index R_i, neighbor similarities $\mathbf{w}_i$ by $f_{\hat{\theta}}$



Method

Feature embeddings of pre-trained CLIP model on ImageNet



“French bulldog” *“Miniature poodle”*

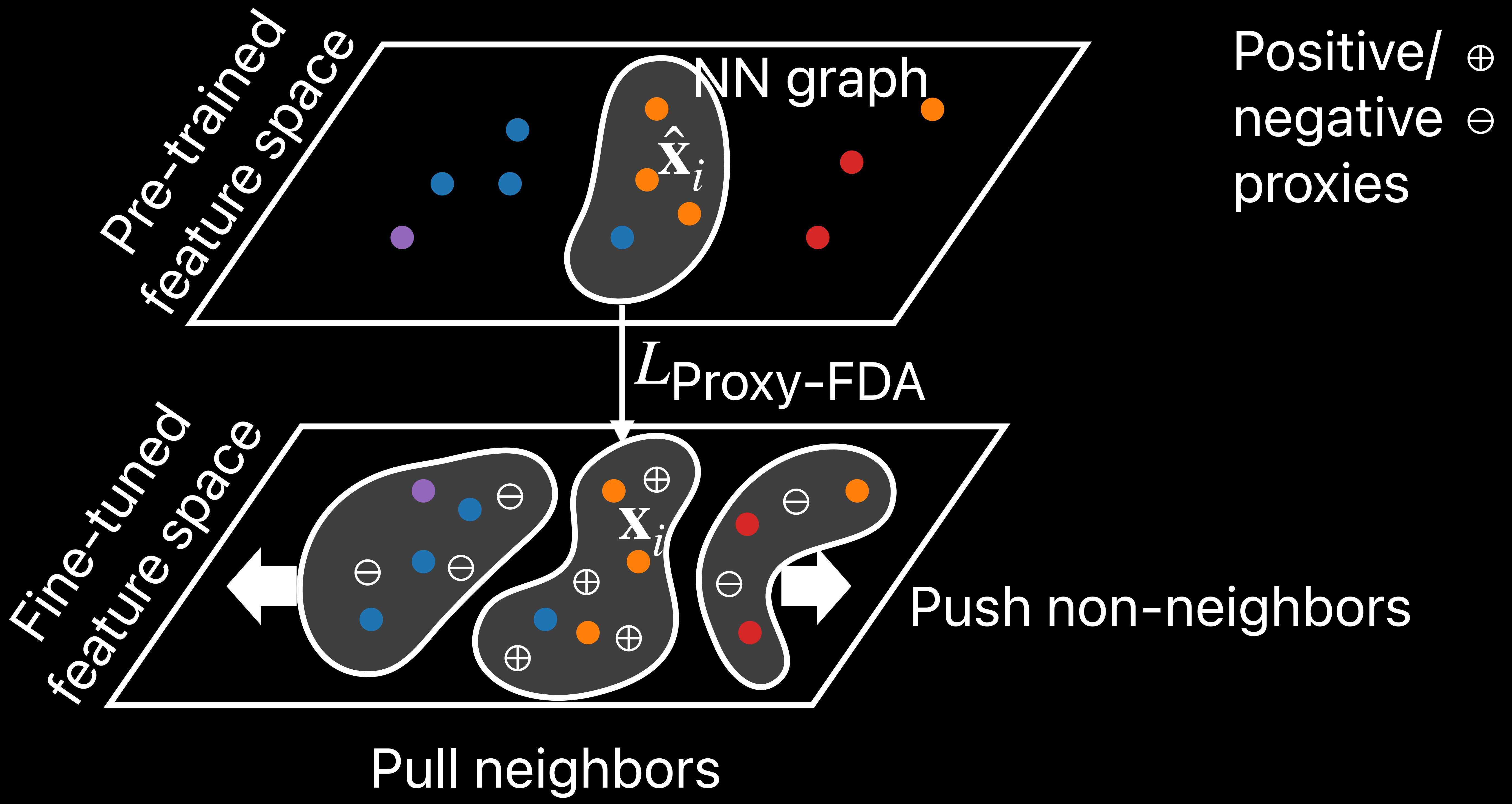
Shared white color attribute

Preserving the common knowledge during fine-tuning maintains the generalizability of foundation models

Method

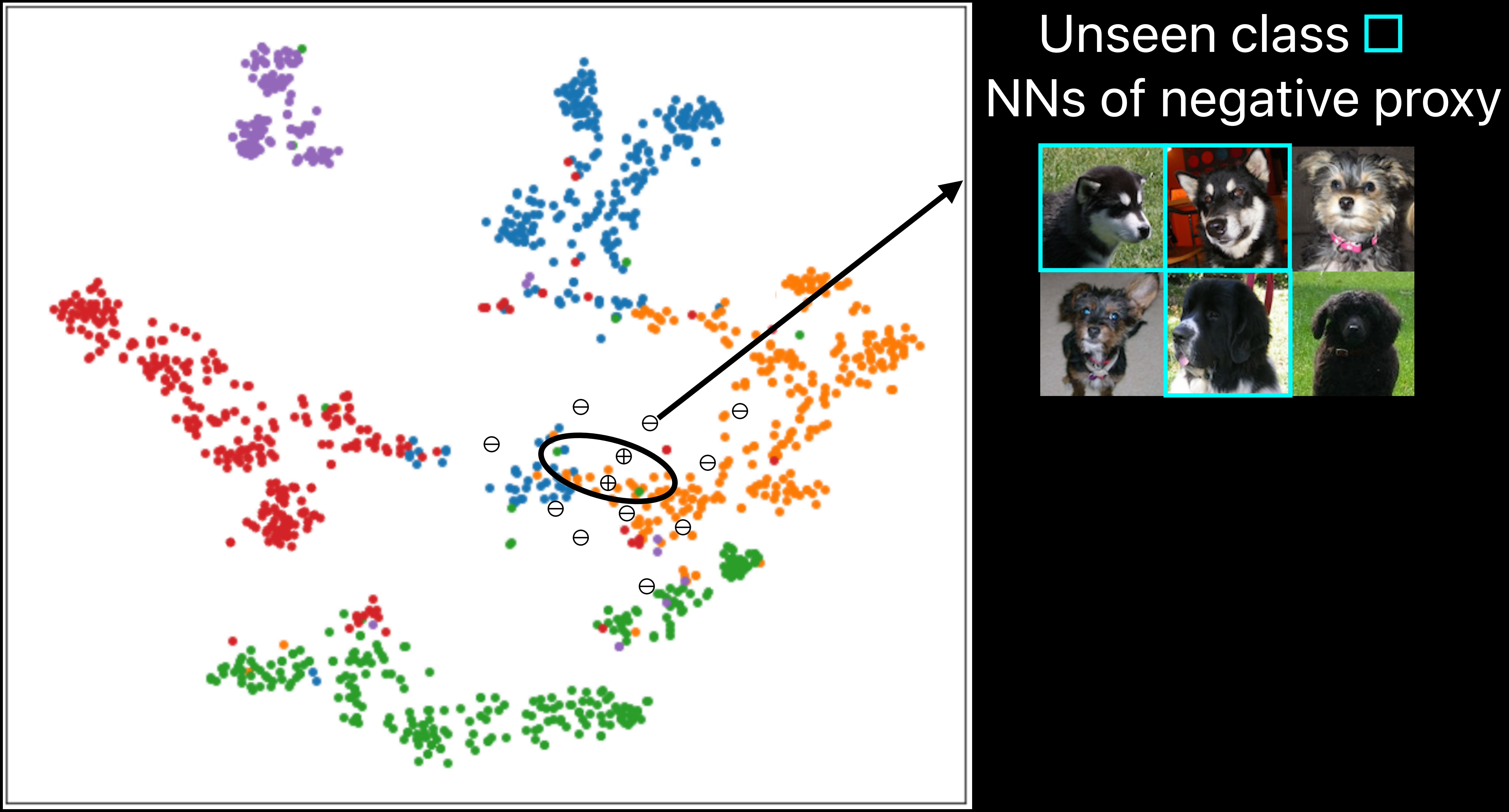
Proxy-FDA

- Helps data-deficient fine-tuning tasks that do not allow sufficient FDA
- Online generator: generate “proxies” as synthetic features to increase diversity



Method

Feature embeddings of pre-trained CLIP model on ImageNet



Proxies improve FDA with richer data/concepts, thereby further reducing concept forgetting

Results

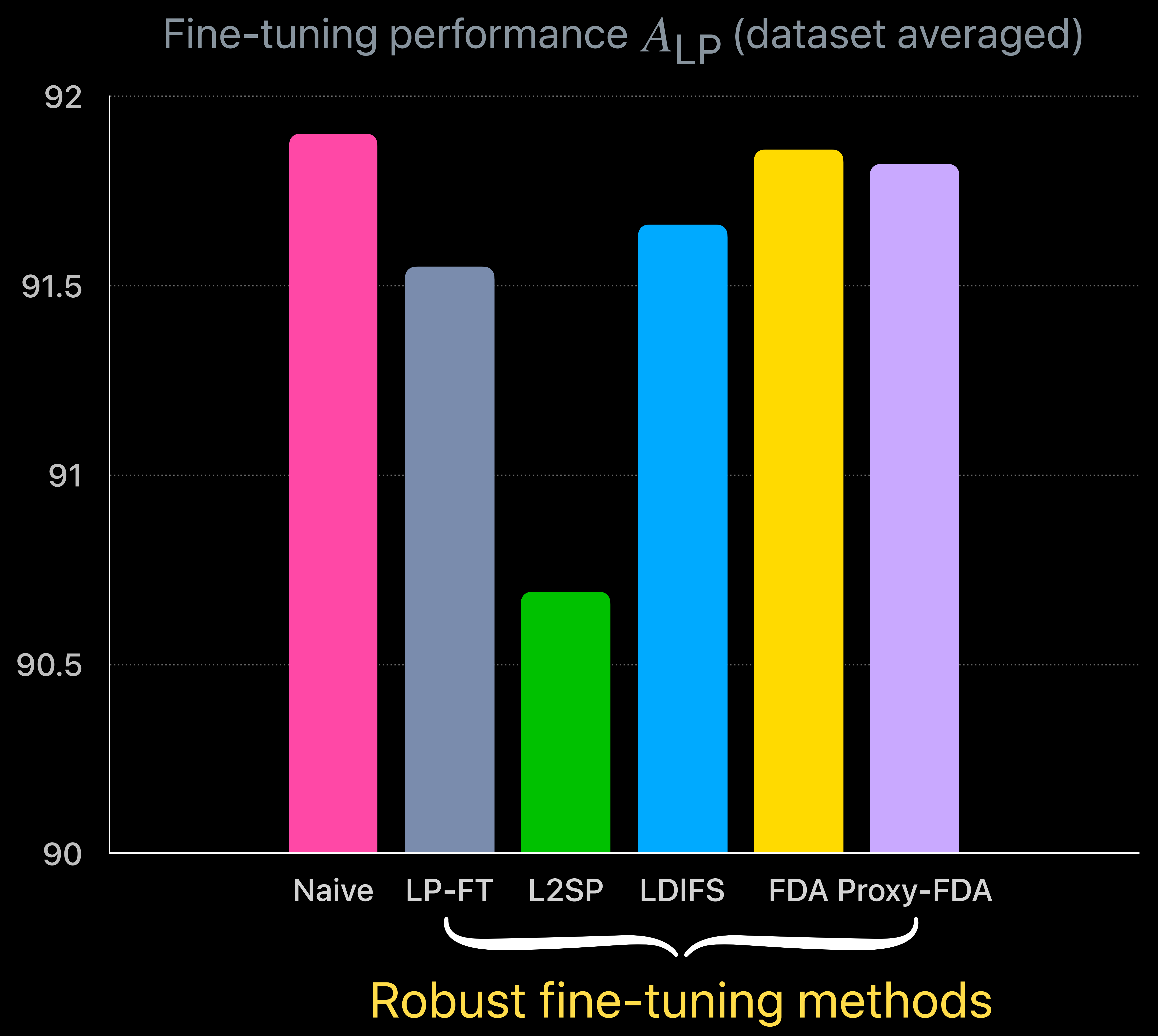
3 fine-tuning settings (classification):

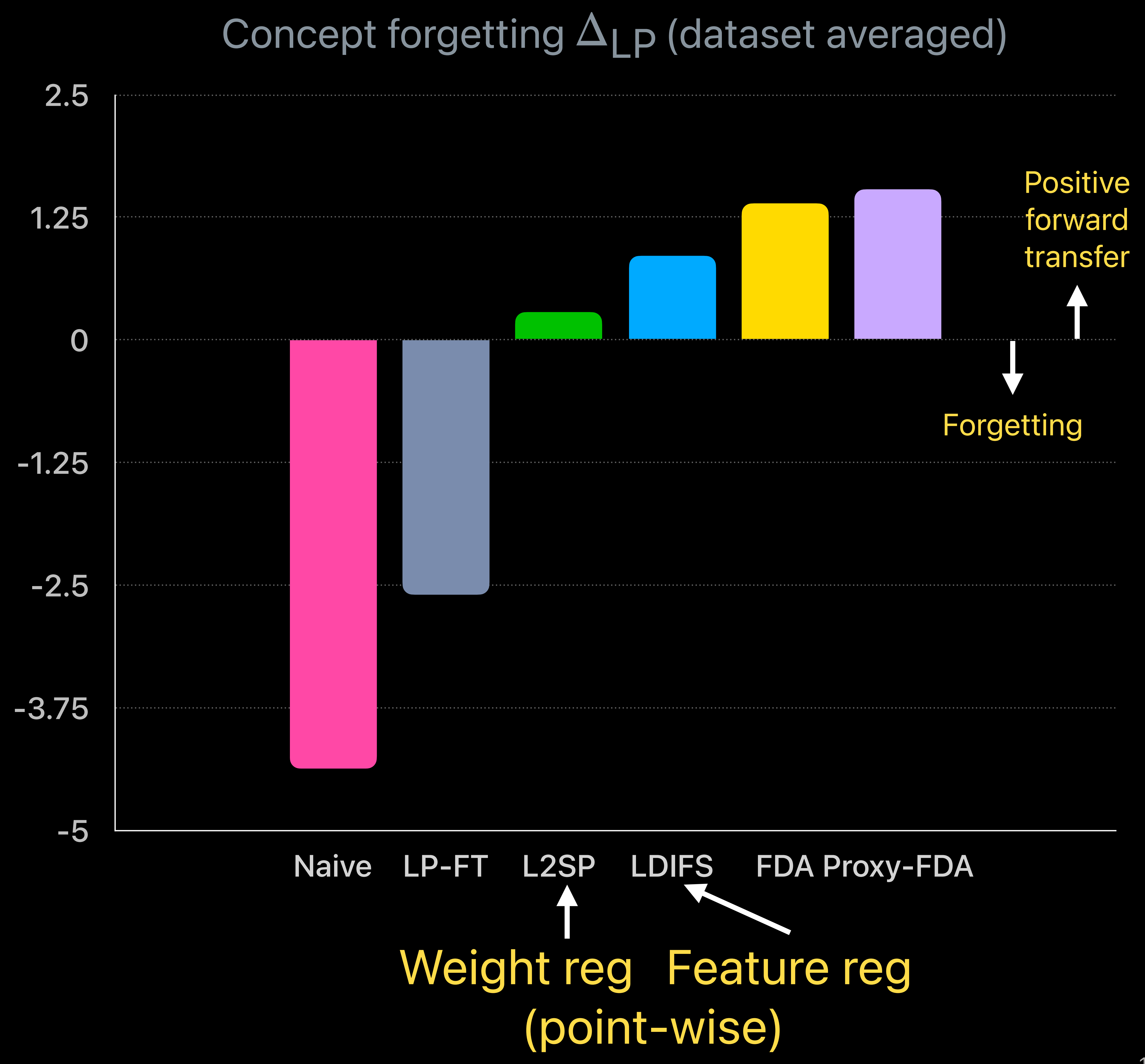
- End-to-end
- Few-shot (more severe forgetting)
- Continual (on a sequence of tasks, catastrophic forgetting)

More fine-tuning tasks beyond classification: captioning, etc

Results

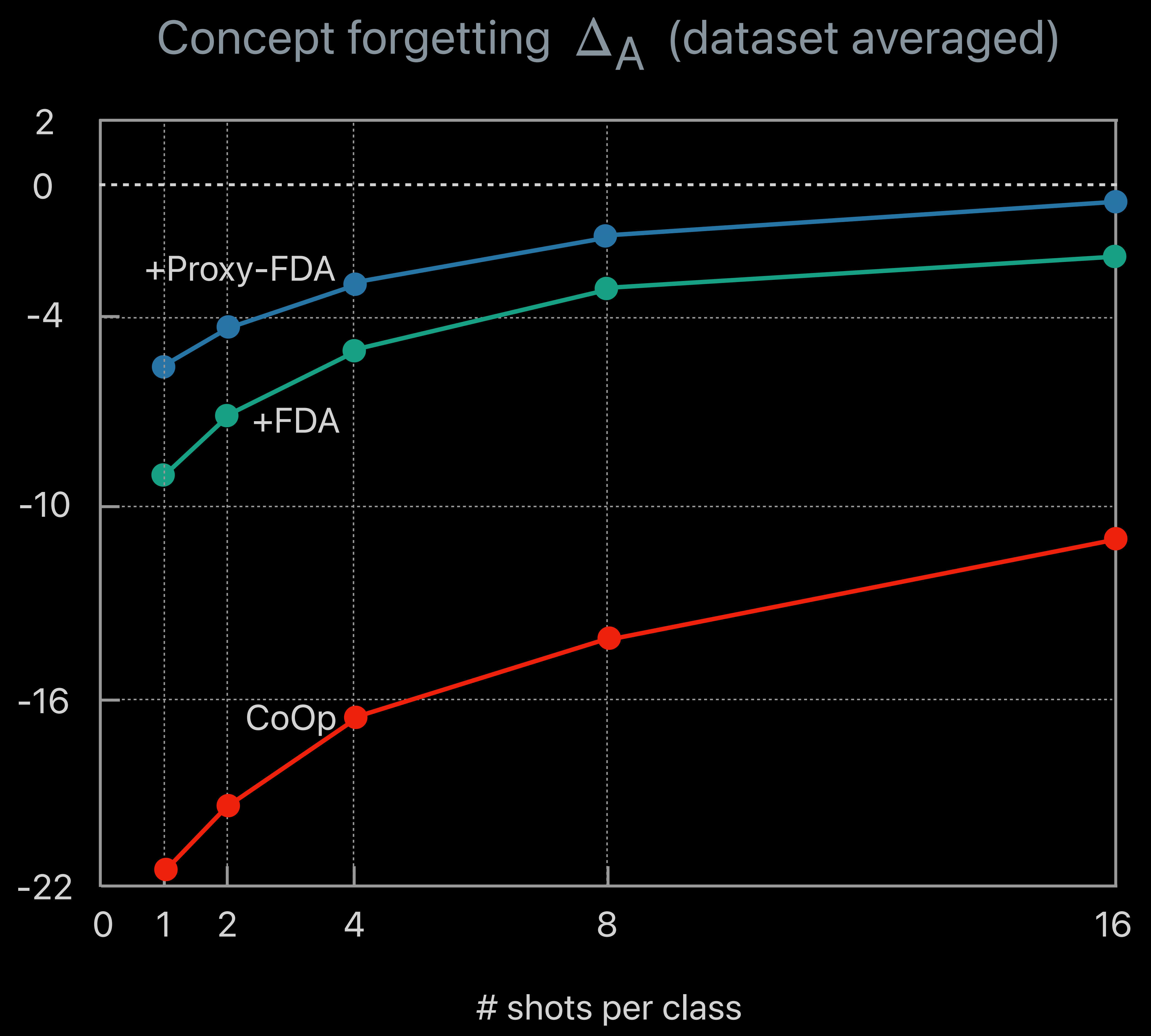
CLIP ViT-B/32: end-to-end fine-tuned on 10 image classification datasets

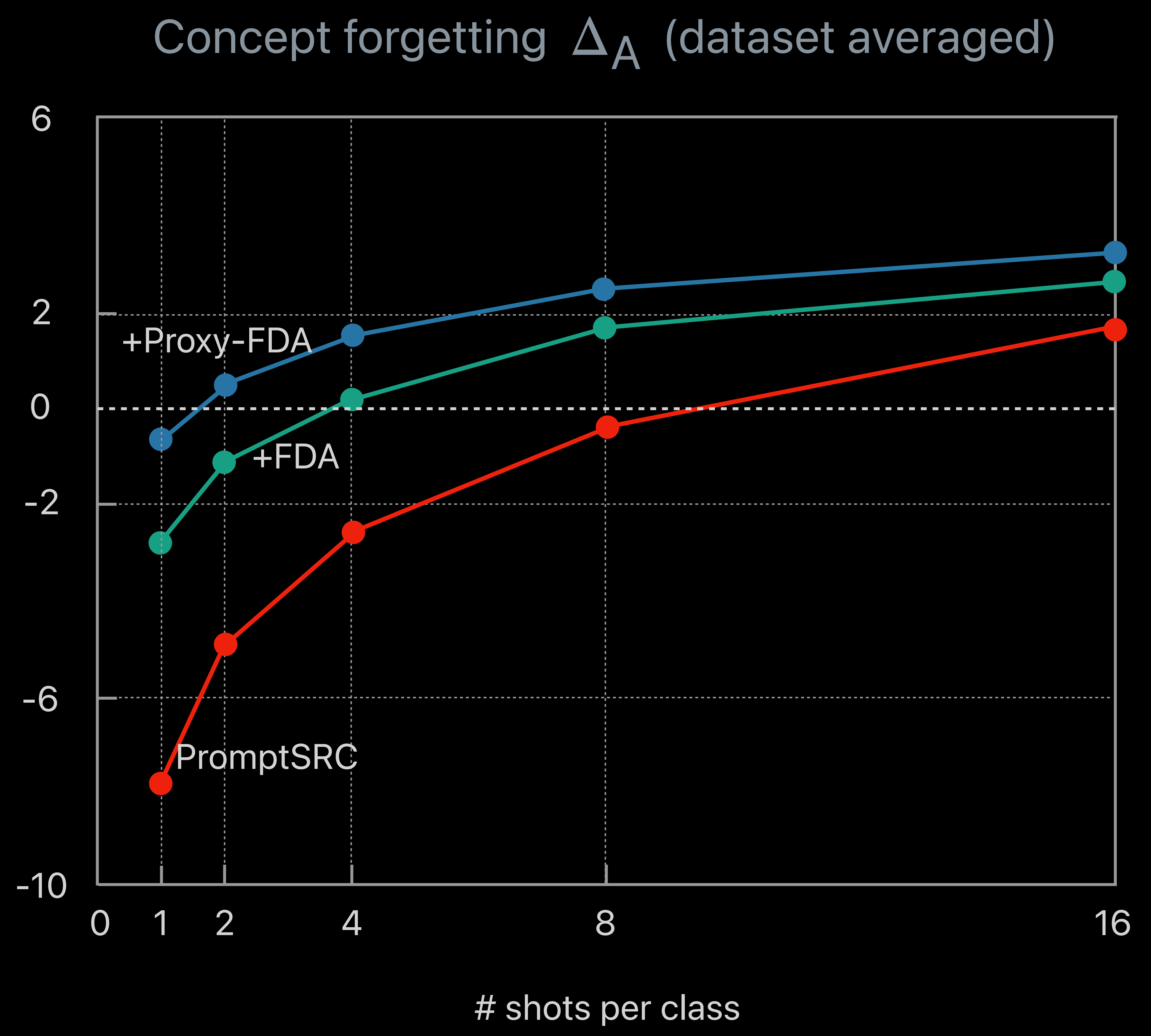




Results

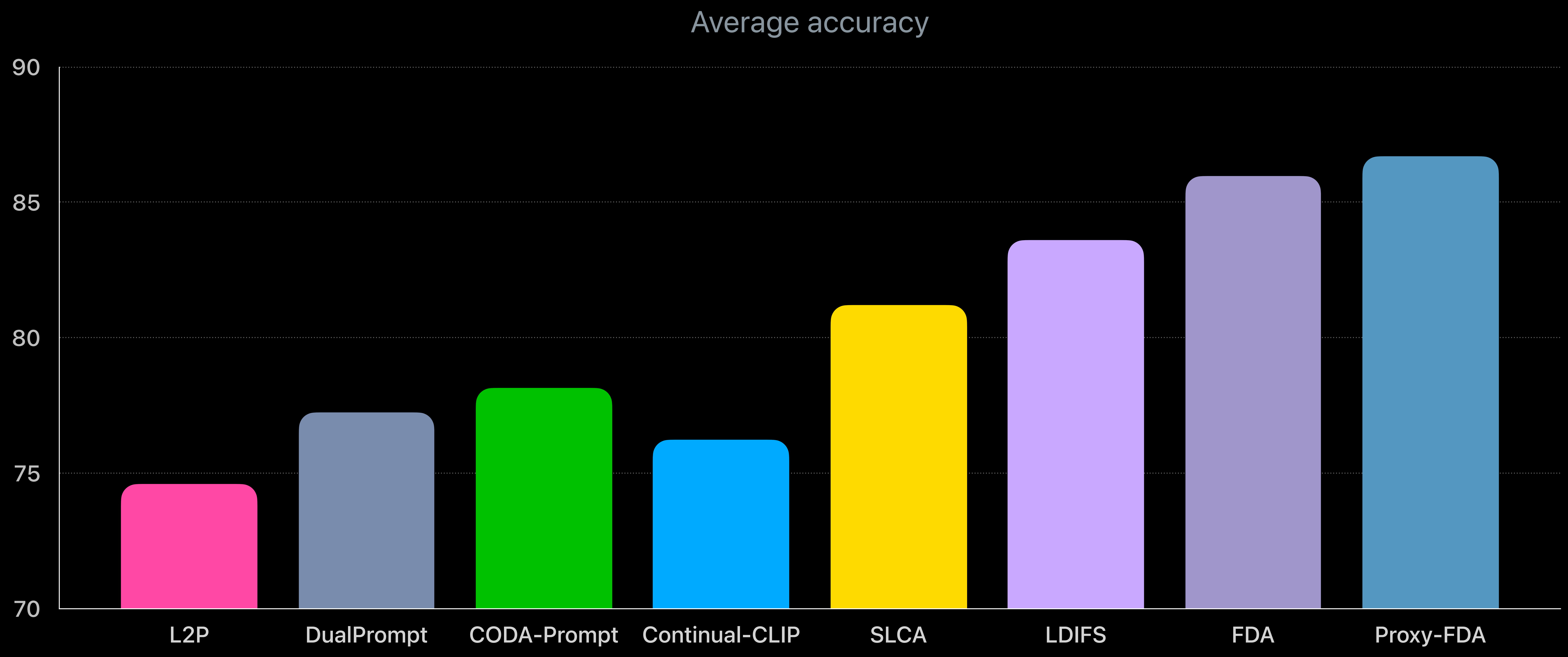
CLIP ViT-B/16: few-shot prompt tuning on 11 image classification datasets





Results

Continual fine-tuning on Split ImageNet-R



Conclusions

- Proxy-FDA preserves concepts when fine-tuning vision foundation models, by aligning feature distribution structures with learned proxies
- State-of-the-art performance on mitigating forgetting in various fine-tuning settings and across different tasks
- Future plan: applications to foundation models beyond vision

TM and © 2025 Apple Inc. All rights reserved.